ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Title goes here

Master Thesis

Robin Bosshard, 16-915-399

October 16, 2023

Supervisors: Prof. Dr. Fernando Perez-Cruz, Dr. Eliza Harris, Lili Gasser (SDSC)
Dr. Kasia Arturi (Eawag)

Department of Computer Science, ETH Zürich

Contents

Chapter 1

Background

This chapter provides background information necessary to understand the rest of the thesis. We introduce the ToxCast's invitro database together with processing pipeline tcpl to get familiar how the bioactivity labels are generated from the in vitro high-throughput screening (HTS) data. We also introduce the concept of molecular fingerprints that are used as the features for the machine learning models.

1.1 The Evolution of Toxicity Testing

Back in 2007, the U.S. National Academy of Sciences introduced a visionary perspective and published a landmark report, titled as *Toxicity Testing in the 21st Century: Vision and Strategy*, advocating a transition from conventional, time-consuming animal-based in vivo tests to efficient high-throughput in vitro pathway assays on cells.

In the realm of high-throughput screening (HTS), a multitude of in vitro bioassays that improve chemical screening can be executed, thanks to the advancements in robotics, data processing and automated analysis. This synergy yields to the generation of extensive toxicity datasets like ToxCast and Tox21.

HTS datasets like ToxCast and other sources opened the door to promising applications of machine learning in predictive computational toxicology. We can develop predictive models to screen environmental chemicals that have little toxicity data available, where the outcomes can be used for further testing prioritization. Such models often forcast toxicity based on chemical structures using Quantitative Structure-Activity Relationships (QSARs) and molecular fingerprints. Molecular fingerprints encode molecules as fixed-length binary vectors, denoting the presence (1) or absence (0) of specific substructures or functional groups.

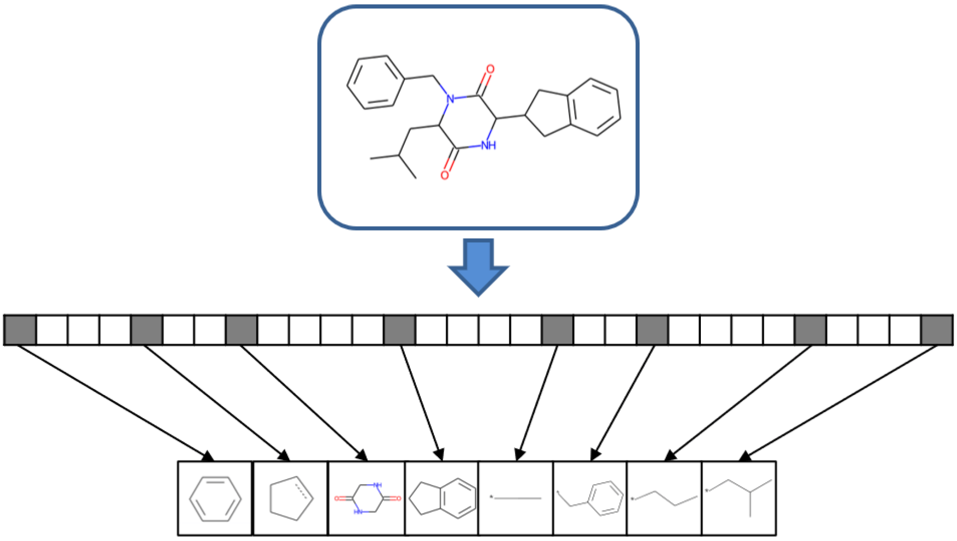


Figure 1.1: Figure 1 adapted from Janela et al (2022) [1]. Schematic molecular fingerprint. Each bit position accounts for the presence or absence of a specific structural fragment. Bit positions are set on (set to 1, gray) if the substructure is present in a molecule, or set off (white) if it is absent.

The utilization of molecular fingerprints for in vitro toxicity prediction is based on the assumption that molecular toxic effects result from relatively straightforward interactions between distinct chemical components and receptors during a molecular initiating event (MEI). On a larger biological scale, the MEI can set in motion a sequential chain of causally linked key events (KE) at different levels of biological organisation from within cells to potentially culminating in an adverse outcome pathway (AOP) at the organ or organism level, as depicted in Figure X. The mechanistic information captured in AOPs reveal how chemicals or stressors cause harm, offering insights into disrupted biological processes, potential intervention points but also guide regulatory decisions on next generation risk assessment and precise toxicity testing. With the AOP framework we have a analytical construct that allows an activity mapping from the presence or absence of certain molecular substructures encoded in the fingerprints to target mechanistic toxicity. Finally, when monitoring disruptions in toxicity pathways, pharmacokinetic models can be leveraged to extrapolate in vitro findings to human blood and tissue concentrations.

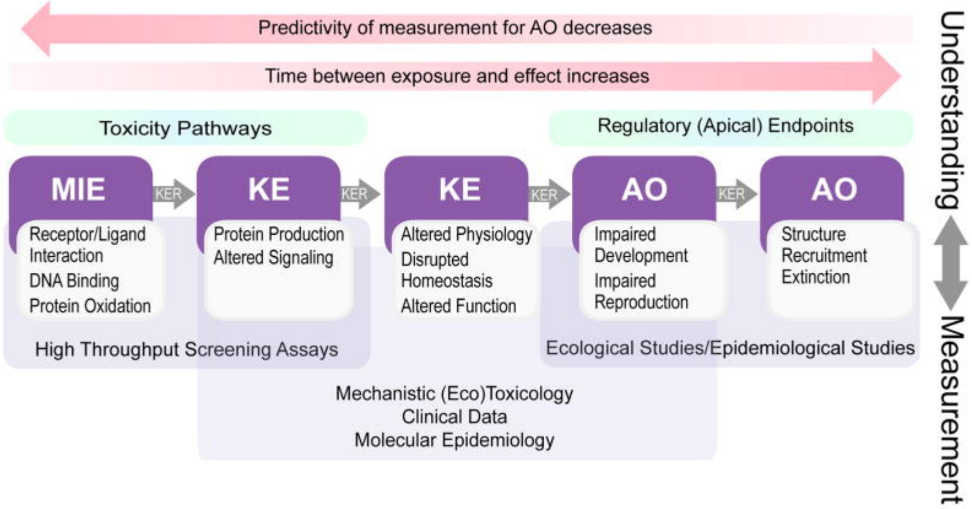


Figure 1.2: Figure 1 from From Ankley and Edwards (2018): Harvesting the promise of AOPs: An assessment and recommendations [2]. Depiction of the role of the adverse outcome pathway (AOP) framework. MIE=Molecular Initiating Event, KE=Key Event, KER=Key Event Relationship, AO=Adverse Outcome

1.2 InvitroDB v4.1

The most recent release of the ToxCast's (Toxicity Forecaster) database [1], referred to as invitroDBv4.1, serves as a source of an extensive collection of high-throughput screening (HTS) targeted bioactivity data. This database encompasses information on a total of 10 196 compounds, selectively screened across 1485 assay endpoints. The assays utilize a range of technologies to assess the impact of chemical exposure on a wide array of biological targets, including individual proteins and cellular processes such as mitochondrial health, developmental processes and nuclear receptor signaling.

This resource originated from the collaboration of two prominent institutions: the United States Environmental Protection Agency[2] (EPA) through its ToxCast program and the National Institutes of Health[3] (NIH) via the Tox21[4] initiative. Utilizing data gathered from various research laboratories, this relational database is publicly available and can be downloaded[5] by visiting the official ToxCast website.

[1] released on September 21, 2023

[2] https://www.epa.gov

[3] https://ntp.niehs.nih.gov/whatwestudy/tox21

[4] https://tox21.gov/

[5] https://www.epa.gov/chemical-research/exploring-toxcast-data

1.3 tcpl v3.0

In Chapter **??**, we introduce the Python re-implementation *pytcpl* of the core components of the ToxCast pipeline *tcpl*, originally an R package. The tcpl package offers a comprehensive suite of tools for managing HTS data, provides consistent and reproducible concentration-response modeling and populates the MYSQL database, invitrodb. The multiple-concentration screening paradigm intends to pinpoint the activity of compounds, while also estimating their efficacy and potency. The concentration-response modeling procedure also addresses outlier robustness and signal loss due to cytotoxicity.

To streamline cross-experiment comparisons and reduce parameter complexity, concentration-response modeling adheres to a zero-centered, positive response paradigm. Negative response data undergoes inverse transformation during normalization. To ensure robustness without data exclusion, a log-likelihood function utilizing a Student's t-distribution with 4 degrees of freedom [3] is employed. The model with the lowest Akaike Information Criterion (AIC) value is selected as the *winning* model. The winning model is then used to estimate the efficacy and potency of the compound. The potency estimates, also called point-of-departure (POD) estimates, are derived from the fitted curve characteristics, identifying concentrations at which the model curve crosses certain response levels. For example, the activity concentration at which the compound reaches 50% of its maximum response is denoted by *ac50* and similarly the activity concentration at cutoff efficacy by *acc*. Addionionally, the package calculates assay noise by computing the median absolute deviation over response values from the first two concentrations (bmad). The baseline region is defined as $0 + 3bmad$, and *acb* is the concentration where the model first reaches 3bmad. The figure illustrates the four POD estimates. To classify[6] a concentration-response series as an active hit and allowing to determine PODs, the following criteria must be met: The *winning* model must be either the Hill or gain-Loss model, with modeled curve's peak surpassing the assay-specific efficacy cutoff, and at least one concentration should have a median response exceeding this threshold. Notably, no POD estimates are computed when the compound is considered inactive[7], as these estimates are not applicable in such cases.

1.3.1 Tcplfit2

To improve upon tcpl, R package *tcplfit2* was developed, a standalone package focused on curve-fitting and hit-calling. The package also offers a more flexible and robust fitting procedure, allowing for the use of different opti-

[6] legacy tcpl employs only binary classification

[7] if the *winning* fit model was the constant model

mization algorithms and the incorporation of user-defined constraints. (Todo: Explain MLE, Compare Strictly standardized mean difference.) The package also includes a more comprehensive set of POD estimates, including the *ac10* and *ac95* estimates. Tcplfit2 differs from other R-language open-source concentration-response packages like *drc* and *mixtox*, as it is specifically tailored for HTS concentration-response data, offering an extensive set of curve models, summarized in Table 1.1.

Table 1.1: tcplfit2 Model Details

Model	**Label**	**Equations**[1]
Constant	cnst	$f(x) = 0$
Linear	poly1	$f(x) = ax$
Quadratic	poly2	$f(x) = a\left(\frac{x}{b} + \left(\frac{x}{b}\right)^2\right)$
Power	pow	$f(x) = ax^p$
Hill	hill	$f(x) = \frac{tp}{1+\left(\frac{ga}{x}\right)^p}$
Gain-Loss	gnls	$f(x) = \frac{tp}{(1+\left(\frac{ga}{x}\right)^p)(1+\left(\frac{x}{la}\right)^q)}$
Exponential 2	exp2	$f(x) = a\left(\exp\left(\frac{x}{b}\right) - 1\right)$
Exponential 3	exp3	$f(x) = a\left(\exp\left(\left(\frac{x}{b}\right)^p\right) - 1\right)$
Exponential 4	exp4	$f(x) = tp\left(1 - 2^{-\frac{x}{ga}}\right)$
Exponential 5	exp5	$f(x) = tp\left(1 - 2^{-\left(\frac{x}{ga}\right)^p}\right)$

[1] Model parameters: a: x-scale, b: y-scale p: (gain) power, q: (loss) power, tp: top, ga: gain AC50, la: loss AC50

All the models assume that the normalized observations derived from the concentration-response series do not conform to a normal distribution. Instead, they adhere to a Student's t-distribution with 4 degrees of freedom. The Student's t-distribution has heavier tails compared to the normal distribution, making it more robust to outlier and eliminates the necessity of removing potential outliers prior to the fitting process. The model fitting algorithm in `tcplFit2` employs maximum likelihood estimation (MLE) to estimate model parameters for all available models.

Consider $t(z, \nu)$ as the Student's t-distribution with ν degrees of freedom, where y_i represents the observed response for the i-th observation, and μ_i is the estimated response for the same observation. The calculation of z_i is as follows:

$$z_i = y_i - \mu_i \exp(\sigma)$$

where σ is the scale term.

Then the log-likelihood is

$$\sum_{i=1}^{n} [\ln(t(z_i, 4)) - \sigma]$$

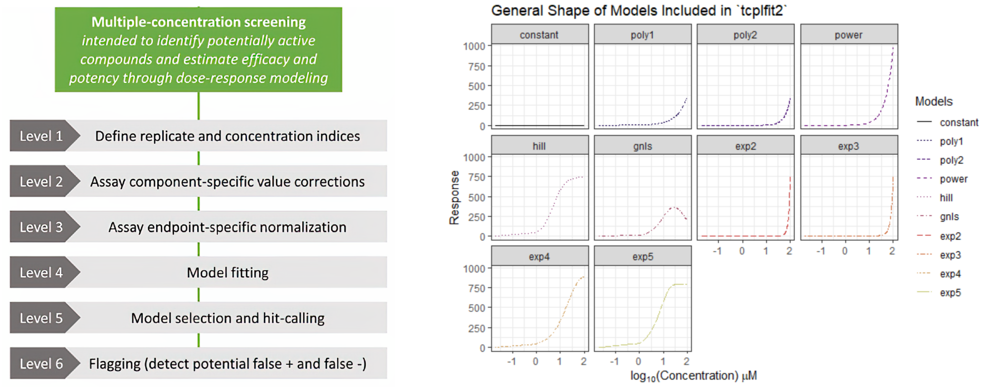
where n is the number of observations.

Hitcalling:

The `tcplhit2_core` continuous hitcall is calculated based on the product of the probabilities of the following values: (i) that at least one median response is greater than the cutoff; (ii) that the top of the fitted curve is above the cutoff, and (iii) that the winning AIC value is less than that of the constant model. The first probability is computed by using the error parameter from the model fit and t-distribution to calculate the odds of at least one response exceeding the cutoff (the error model around the data uses a 3-parameter t-distribution). The second is by using the likelihood ratio to compute the one-sided probability of the cutoff being exceeded. The third is set to be the Akaike weight relative to the constant model:

$$\frac{e^{-\frac{1}{2}AIC_{\text{winning}}}}{e}^{-\frac{1}{2}AIC_{\text{winning}}+e^{-\frac{1}{2}AIC_{\text{cnst}}}}.$$

Multiple-concentration processing includes six processing levels. Briefly, level 1 processing defines concentration and replicate indices, giving integer values 1. . . N to increasing concentrations and technical replicates, where 1 represents the lowest concentration or first technical replicate. Level 2 processing allows for basic transformations of the raw data, e.g., logarithmic conversion, and removes data deemed poor quality by the user. Similar to level 1 in single-concentration processing, level 3 normalizes data to fold-change or percent-of-control and converts concentrations to a logarithm scale. At level 4 data are modeled (described below), before level 5 processing defines the winning model and the activity call. Level 5 processing also separates each data series into categories to facilitate easy triaging of the results. Level 6 processing identifies potential false positive and false negative results, giving problematic data series a flag.


Multiple-concentration screening
intended to identify potentially active compounds and estimate efficacy and potency through dose-response modeling
Level 1 Define replicate and concentration indices
Level 2 Assay component-specific value corrections
Level 3 Assay endpoint-specific normalization
Level 4 Model fitting
Level 5 Model selection and hit-calling
Level 6 Flagging (detect potential false + and false -)
General Shape of Models Included in `tcplfit2`
constant
poly1
poly2
power
hill
gnls
exp2
exp3
exp4
exp5
Response
1000
750
500
250
0
-1
0
1
2
log$_{10}$(Concentration) μM
Models
constant
poly1
poly2
power
hill
gnls
exp2
exp3
exp4
exp5

(a) Overview of tcpl's data analysis pipeline for multiple concentrations screening data.

(b) Employed curve-fit models in tcpl v3.0 for fitting concentration-response data series through the application of maximum likelihood estimation.

Figure 1.3: tcpl

1.4 Chemical Target Toxicity vs. Cytotoxicity

Intuitively, we expect increasing chemical concentrations to result in increasing bioactivity. However, this is not always the case. At higher doses the chemical can become cytotoxic leading to dying cells. In consequently, a reduction in bioactivity can occur, e.g., the activation of the reporter gene decreases.

Chemical toxicity can manifest in diverse ways, falling into two major categories [4]:

1. **Specific toxicity** is the result of a chemical's interaction and disruption of a specific biomolecular target or pathway, such as a receptor agonist/antagonist effect or enzyme activation/inhibition.
2. the **Cytotoxicity and cell stress** is the generalized disruption of the cellular machinery. Cell-disruptive processes encompass various mechanisms, such as protein, DNA, or lipid reactivity, or processes like apoptosis, oxidative stress responses or mitochondrial disruption. Cell viability can be evaluated either individually or concurrently. One approach is to assess it by calculating the proportion of live cells in a population, employing a fluorescent dye that specifically enters living cells. This dye remains incapable of permeating the membranes of deceased cells, resulting in fluorescence intensity directly correlating with cell viability.

It is common for compounds to exhibit target bioactivity within a limited concentration range, which coincides with a nonspecific activation response in the presence of cell stress and cytotoxicity. Figure 1.4 illustrates the interference between specific toxicity and cytotoxicity.

A related phenomenon is referred to as the *cytotoxicity burst* [4], where e.g., the reporter genes can be non-specifically induced close to cell death [5]. The Toxcast pipeline is designed to minimize false negatives due to the inclusive risk assessment.Nevertheless, attributed to interference processes, the reliability of reported activities becomes uncertain and false positives are possible without a cytotoxicity evaluation. While a portion of the assay activity within this concentration range might indeed reflect chemical interactions with the intended assay target, another portion does not.

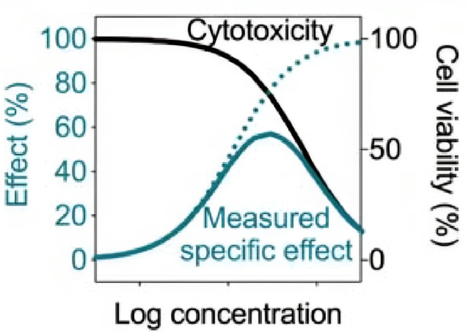


Figure 1.4: Figure 7.8 from Escher at al. [5]: Bioanalytical Tools in Water Quality Assessment: Second Edition. Example of a bioassay response with cytotoxicity interference. The dotted line shows the theoretical effect but due to cytotoxicity (black line is cell viability), the measured effect has an inverted U-shape. The measured effect can addionally be confounded and intensified by the cytotoxicity burst, where even an exponential shape is likely for the gaining part. In this case, the effect should be only evaluated up to some concentration.

1.5 Challenges of HTS data

1.6 Molecular Fingerprints

Bibliography

[1] T. Janela, K. Takeuchi, and J. Bajorath, "Introducing a chemically intuitive core-substituent fingerprint designed to explore structural requirements for effective similarity searching and machine learning," *Molecules*, vol. 27, no. 7, 2022, ISSN: 1420-3049. DOI: 10.3390/molecules27072331. [Online]. Available: https://www.mdpi.com/1420-3049/27/7/2331.

[2] G. T. Ankley and S. W. Edwards, "The adverse outcome pathway: A multifaceted framework supporting 21st century toxicology," *Current Opinion in Toxicology*, vol. 9, pp. 1–7, 2018, Risk assessment in Toxicology, ISSN: 2468-2020. DOI: https://doi.org/10.1016/j.cotox.2018.03.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468202017301420.

[3] "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2290063 (visited on 09/20/2023).

[4] R. Judson *et al.*, "Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space," *Toxicological Sciences*, vol. 152, no. 2, pp. 323–339, May 2016, ISSN: 1096-6080. DOI: 10.1093/toxsci/kfw092. eprint: https://academic.oup.com/toxsci/article-pdf/152/2/323/26290632/kfw092.pdf. [Online]. Available: https://doi.org/10.1093/toxsci/kfw092.

[5] B. Escher, P. Neale, and F. Leusch, *Bioanalytical Tools in Water Quality Assessment*. IWA Publishing, Jun. 2021, ISBN: 9781789061987. DOI: 10.2166/9781789061987. eprint: https://iwaponline.com/book-pdf/899726/wio9781789061987.pdf. [Online]. Available: https://doi.org/10.2166/9781789061987.